VTT Technical Research Centre of Finland

Origin-destination matrix estimation with a conditionally binomial model

Kuusela, Pirkko; Norros, Ilkka; Kilpi, Jorma; Räty, Tomi

Published in:
European Transport Research Review

DOI:
10.1186/s12544-020-00433-7

Published: 17/06/2020

Document Version
Publisher's final version

European Transport Research Review

ORIGINAL PAPER Open Access

Origin-destination matrix estimation with a conditionally binomial model

Pirkko Kuusela[1]*, Ilkka Norros[2], Jorma Kilpi[1] and Tomi Räty[1]

Abstract

A doubly stochastic, conditionally binomial model is proposed to describe volumes of vehicular origin-destination flows in regular vehicular traffic, such as morning rush hours. The statistical properties of this model are motivated by the data obtained from inductive loop traffic counts. The model parameters can be expressed as rational functions of the first and second order moments of the observed link counts. Challenges arising from the inaccuracy of moment estimates are studied. A real origin-destination traffic problem of Tampere city is solved by optimisation methods and the accuracy of the solution is examined.

Keywords: Vehicular traffic, Origin-destination flow, Traffic matrix, Doubly stochastic, Method of moments, Optimisation

1 Introduction

Origin-destination (O-D) matrices are required for many transport modelling and planning purposes, in particular, if there are changes in the use of land, economic states, or transportation habits. Consider a connected non-complete graph, and assume that a non-negative flow with a fixed route is associated to each ordered pair of nodes. The matrix of the flow volumes is often called the traffic matrix or the O-D matrix of the system. Typically, the traffic matrix cannot be inferred from the aggregated flows observed on edges, as the number of edges is smaller than the number of node pairs. Since the O-D volumes in a transportation network are valuable information in many contexts and the available observations often consist only on edge flow measurements (e.g., vehicle counts on a road section), traffic matrix estimation has been an interest for decades.

The O-D estimation problems need to be studied from observability and identifiability points of view. Observability holds, when short-term O-D flows can be uniquely inferred from the observations, see [3, 5]. A recent work on observability involves, e.g., problems of designing measurements and input control [6, 23]. The notion of identifiability refers to the uniqueness of parameters that govern the underlying probability distributions of travel demands. The concepts of observability, identifiability and estimation are explained in a unifying manner by [25].

When the data consist of a time series of edge count vectors, the system is seldom observable, but possibly statistically identifiable. The O-D matrix estimation has a long history — see, e.g., the review by [1]. We point to the seminal paper [21] on identifiability of the O-D matrix and work in [12] on vehicular O-D matrix estimation with variety of stochastic models and statistical methods. [20] continued [21] with identifiability theorems for wider classes of distributions, still assuming, however, statistical independence of basic blocks of traffic.

Traffic random variables and models have been covered from the probabilistic and physical point of view by [4]. O-D flows have typically been modelled using Poisson, multinomial, uniform, gamma or Gaussian distributions. The use of doubly stochastic models for road traffic has been quite rare. However, [17] analysed a doubly stochastic model, where a stationary process runs on the day timescale. [14] compares Poisson and doubly stochastic negative binomial models. Different Poisson mixtures and

*Correspondence: pirkko.kuusela@vtt
[1]Techical Research Centre af Finland, VTT Ltd., P.O. Box 1000, 02044 VTT, Espoo, Finland
Full list of author information is available at the end of the article

two other hierarchical structures are examined in [18] with a Bayesian estimation approach.

Optimisation offers a general methodology for solving observability, estimation and prediction problems in traffic networks. A review by [6] examines these problems in a unified framework. Recently, [24] has generalized method of moments through integrating statistics and optimisation techniques with transportation domain knowledge.

This study was motivated by the observation that traffic counts were positively correlated also between the edges that had negligible physical connection in terms of traffic flows, e.g., between the opposite directions of one street section. This calls for models, in which some common factors influence the traffic variation from day to day. The correlations caused by shared flows are then shadowed by those caused by the common factors, and the latter have to be filtered out to facilitate utilisation of the former. On the other hand, it is natural to assume that the random variations of distinct O-D flows are conditionally independent, given the common factors. Thus, the observed positive correlations prompt to consider doubly stochastic models. Such models suit well to European cities that offer multiple ways to commute and people favour the most suitable one depending on some common factor, e.g., weather.

Our traffic matrix estimation approach is similar to [12], but with a conditionally binomial model of vehicular traffic. To our knowledge, such a model has not been studied before in this context. Our graph structure is the simplest interesting case that illustrates the problem of traffic splitting between two paths. However, our topology model can also be seen as a network abstraction as discussed in [12]. Our vehicle count data suggested modelling the traffic by a conditionally binomial distribution. The model is too simple to be considered as an accurate representation of the reality, but it captures those features we consider to be the most essential. However, the conditionally binomial model exposes us to statistical challenges illustrated in the paper.

The results in this paper are threefold. First,we show that a doubly stochastic model agrees with statistical properties of our data and that the model parameters can be computed as rational functions of the first and second order moments of the observed edge counts. Second, the estimation turns out to be practically feasible only when the overall vehicle population is not too large. This problem can, however, be alleviated by optimisation. Third, experiments with such techniques are made using real world data, focusing on rush hours of working day mornings. The overall conclusion is that a doubly stochastic structure can be considered adequate for regular rush hour traffic at least, but in our case the model parameters lie in an area in which only very rough estimation is possible.

Our network topology and the conditionally binomial traffic model is presented in section 2, followed by statistical methodology in section 3. In section 4, the model and the optimisation methods are tested based on traffic data from Tampere, Finland. Conclusions are drawn in section 5.

2 Modelling framework

The modelling framework is discussed mainly in the context of a simple model, in which the theoretical development and the real traffic case study illustrate the main characteristics of the framework. However, we elaborate also a more complex topology in section 2.3 and derive equations that apply to any network topology and route system.

2.1 Minicity model

We focus on a simple network abstraction to study how well the amount of local and through traffic can be inferred from minimum measurements of the total traffic. Our network, called "Minicity", is shown in Fig. 1.

There are 3 sink/source nodes, denoted by W=West, C = Centre, E=East, and 4 measurement points, denoted by 1, 2, 3, 4. There are 6 flows with O-D pairs $(W, C), (W, E), (C, E), (E, C), (E, W)$, and (C, W). At each measurement point, we observe the number of vehicles during fixed time intervals. In the West-East direction, denote the traffic flows $F_{(W,C)}, F_{(C,E)}, F_{(W,E)}$ as X, Y, Z, and the observations as $O_1 = X + Z$, $O_2 = Y + Z$. The Minicity model is not observable (see [5] and [25]): the flows X, Y, Z and $X - \delta, Y - \delta, Z + \delta$, with δ some constant, yield identical observations O_1 and O_2. However, the analysis is possible as a stochastic demand estimation problem, in which a traffic model has an important role in identifiability.

2.2 Stochastic modelling of regular traffic flows

We focus on the flow structure of regular traffic consisting of a rather stable amount of vehicles. Assume that

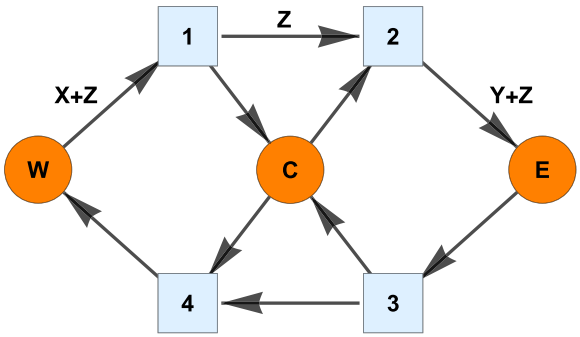

Fig. 1 The Minicity model

the observation data consist of a long sequence of measurements on usual working days. Then the traffic follows steady patterns at large, but stochastic variation makes each day a bit different. To get an idea of this variation, Fig. 2 presents the variation of two very illustrative quantities on working day Mondays at the resolution of 15 minutes. The upper plot shows the index of dispersion of counts (IDC; variance over mean) and the lower plot the coefficient of variation (CoV; standard deviation over mean).

First, recall that the IDC of a Poisson distribution always equals one. The upper plot shows that the traffic can be considered Poissonian during the most silent period of the night, but not at any other time. However, the IDC is relatively low (about 2) and stable also during the day between 9:30 and 15:30, whereas it is notably higher during the morning and afternoon rush hours. The non-Poissonian character of vehicular traffic is generally recognised, but the upper plot of Fig. 2 offers a more detailed picture.

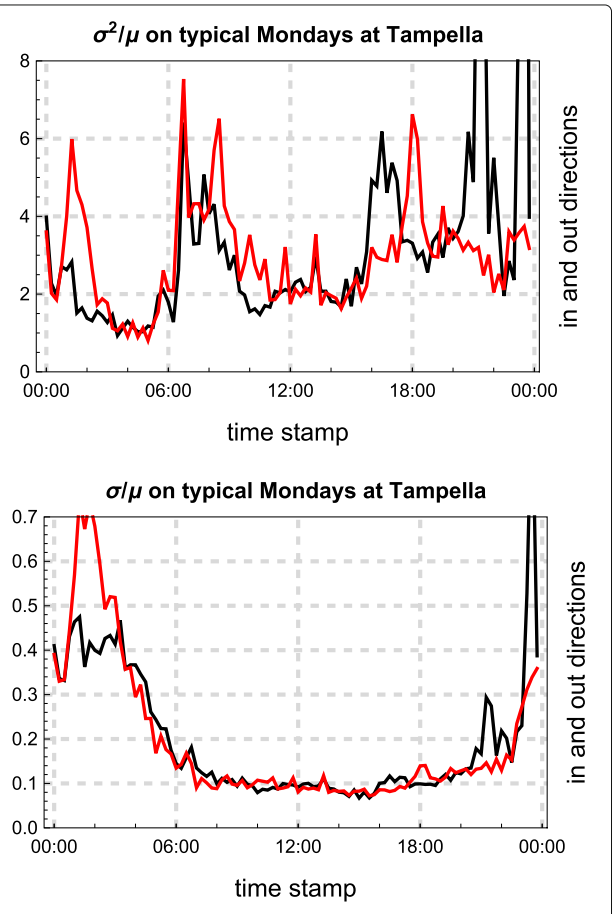

Fig. 2 The index of dispersion (IDC, σ^2/μ,upper) and the coefficient of variation (CoV, σ/μ, lower) of the observed vehicle counts are illustrated on normal working day Mondays at Tampella measurement point. Inbound and outbound traffic are illustrated in black and red, respectively

Although a popular extension of Poissonian modelling has been to introduce additional variables, ϵ and θ, to scale the relation between the variance and the mean, e.g., $\mathrm{Var}(X) = \epsilon(\mathrm{E}(X))^{\theta}$ [7], we study a more specific model.

In a normal working day, the majority of the morning rush hour traffic consists of people's regular drives to workplaces, educational institutions etc. There is a steady population of the same vehicles that create this traffic, but the actual presence of each particular vehicle can be considered as random. This suggests a binomial type of model; from n vehicles potentially appearing in a rush time interval I, each one actually appears with some probability p. Such a model is, however, immediately refuted by statistical data. The IDC of $Bin(n,p)$ is $1-p$, i.e., still lower than that of a Poisson distribution. Obvious shortcomings in the motivation of the above binomial model are that the probability of using a car (i) is not homogeneous in the population, and (ii) varies according to time by common external factors like weather, seasonal holiday activity, flu epidemics etc. The inhomogeneity challenge is serious, but it is not in the scope of this paper. Instead, we focus on the time variation by considering a doubly stochastic, conditionally binomial model, in which the parameter p is the same for all flows but varies from day to day as a stochastic process.

The presence of a common "activity factor" of a day is suggested by the data; we observed a strong positive correlation of measured vehicle counts at separate locations that cannot logically contain more than a negligible number of the same cars, i.e., at the opposite directions of the same road. For a comprehensive statistical analysis of the data utilised in this paper, see [16]. In particular, the correlation matrices of the measurement locations $O_1, \ldots, O_4$ on Monday – Thursday mornings at 7:45 – 8:45 `A.M.` and 8:45 – 9:45 `A.M.`,

	O_1	O_2	O_3	O_4
O_1	1	0.87	0.55	0.50
O_2	0.87	1	0.60	0.51
O_3	0.55	0.6	1	0.79
O_4	0.50	0.51	0.79	1

and

	O_1	O_2	O_3	O_4
O_1	1	0.94	0.74	0.56
O_2	0.94	1	0.77	0.58
O_3	0.74	0.77	1	0.80
O_4	0.56	0.58	0.80	1

,

show the highest correlations between physically connected locations (motorway passing the city). However, *all* location pairs indicate considerable positive correlation. We interpret this common correlatedness to be due to a kind of "activity factor", see also [2].

The assumption of an "activity factor" is also compatible with the lower plot of Fig. 2. The CoV is roughly constant throughout the working time, the rush hours included, despite of strongly differing traffic volumes. This means that the random variation of the daily sequence of traffic counts in a fixed quarter-hour is dominated by a common random factor, not by individual level variation, whose aggregated variance scales with n, not n^2. A detailed analysis of the variance in the conditionally binomial model is given in the next subsections.

2.3 Conditionally binomial model

Consider a fixed time interval I of a day, and the daily traffic counts in I at four measurement stations as depicted in Fig. 1. Assume that for each day k, $k = 1, 2, \ldots$, there is a random variable γ_k with values in $(0, 1)$, interpreted as the "activity level" of day k. The sequence (γ_k) is assumed stationary and ergodic, i.e., its time average agrees with its average over the probability space. Let us then assume that, conditioned on γ_k, the daily West-East traffic flows in period I, denoted as X_k^{WE}, Y_k^{WE}, and Z_k^{WE} (see section 2.1) are independent and have distributions $Bin(n_X^{WE}, \gamma_k)$, $Bin(n_Y^{WE}, \gamma_k)$, $Bin(n_Z^{WE}, \gamma_k)$, respectively, and similarly for the East-West counterparts. Assume Minicity to be uncongested so that the traffic via the city centre is not a realistic option for the through traffic and there is only one path for each O-D flow (see [11] for estimation under congestion).

To simplify, we use the notation $n_X = n_X^{WE}, n_Y = n_Y^{WE}, n_Z = n_Z^{WE}, m_X = n_X^{EW}$, $m_Y = n_Y^{EW}$, $m_Z = n_Z^{EW}$. When considering the model of a single day, the index k is omitted. The n and m parameters represent the sizes of relatively stable populations of vehicles that can be used to make a trip from an origin to a destination, hence passing through some of the measurement points. The numbers $n_X, n_Y, n_Z, m_X, m_Y, m_Z$ as well as the distribution of γ are considered as unknown.

One obvious candidate for modelling γ is the family of beta distributions, resulting in the so-called beta-binomial distributions. In a beta-binomial distribution with parameters (α, β, n) with large n, the squared CoV is approximately β/α, i.e. independent of n, which coincides well with the properties of the measured traffic.

Consider the measured quantities $O_1 = X + Z$ and $O_2 = Y + Z$, and denote their first and second moments as $m_1 = \mathrm{E}(O_1)$, $m_2 = \mathrm{E}(O_2)$, $v_1 = \mathrm{Var}(O_1)$, $v_2 = \mathrm{Var}(O_2)$ and $c_{12} = \mathrm{Cov}(O_1, O_2)$. Using the equations

$$\begin{aligned} \mathrm{Var}(U) &= \mathrm{E}(\mathrm{Var}[U \mid \mathcal{A}]) + \mathrm{Var}(\mathrm{E}[U \mid \mathcal{A}]), \\ \mathrm{Cov}(U, V) &= \mathrm{E}(\mathrm{Cov}[U, V \mid \mathcal{A}]) \\ &\quad + \mathrm{Cov}(\mathrm{E}[U \mid \mathcal{A}], \mathrm{E}[V \mid \mathcal{A}]) \end{aligned} \tag{1}$$

that hold for any square integrable U, V and any conditioning random variable $\mathcal{A}$, we obtain the following expressions in the conditionally binomial model:

$$\begin{aligned} m_1 &= (n_X + n_Z)\mathrm{E}(\gamma), \\ m_2 &= (n_Y + n_Z)\mathrm{E}(\gamma), \\ v_1 &= (n_X + n_Z)^2\mathrm{Var}(\gamma) \\ &\quad + (n_X + n_Z)(\mathrm{E}(\gamma) - (\mathrm{E}(\gamma))^2 - \mathrm{Var}(\gamma)), \\ v_2 &= (n_Y + n_Z)^2\mathrm{Var}(\gamma) \\ &\quad + (n_Y + n_Z)(\mathrm{E}(\gamma) - (\mathrm{E}(\gamma))^2 - \mathrm{Var}(\gamma)), \\ c_{12} &= (n_X + n_Z)(n_Y + n_Z)\mathrm{Var}(\gamma) \\ &\quad + n_Z(\mathrm{E}(\gamma) - (\mathrm{E}(\gamma))^2 - \mathrm{Var}(\gamma)). \end{aligned} \tag{2}$$

Corresponding equations hold for the East-West direction, we denote them by $(2)^{EW}$. The role of covariance c_{12} is essential in identifying n_Z as the other equations involve either $n_X + n_Y$ or $n_Y + n_Z$. Note that because the value of γ is common for both directions, (2) and $(2)^{EW}$ together provide 10 equations for 8 unknowns: $n_X, n_Y, n_Z, m_X, m_Y, m_Z, \mathrm{E}(\gamma)$, and $\mathrm{Var}(\gamma)$.

Let us denote the IDCs and squared CoVs of O_1 and O_2, respectively, as

$$\delta_1 = \frac{\mathrm{Var}(O_1)}{\mathrm{E}(O_1)} = \frac{v_1}{m_1}, \quad \delta_2 = \frac{\mathrm{Var}(O_2)}{\mathrm{E}(O_2)} = \frac{v_2}{m_2},$$
$$\zeta_1 = \frac{\mathrm{Var}(O_1)}{(\mathrm{E}(O_1))^2} = \frac{v_1}{m_1^2}, \quad \zeta_2 = \frac{\mathrm{Var}(O_2)}{(\mathrm{E}(O_2))^2} = \frac{v_2}{m_2^2}.$$

Now, straightforward computations yield the following result.

Proposition 1 *Assume that the moment Eqs. 2 hold for some* $\mathrm{E}(\gamma) \in (0, 1)$, $Var(\gamma) \in [0, \mathrm{E}(\gamma)(1 - \mathrm{E}(\gamma))]$ *and* $n_X, n_Y, n_Z \geq 0$.

1. *When* $n_X \neq n_Y$ *and* $Var(\gamma) > 0$, *(2) has a unique solution whose components are rational functions of the moments and can be written as*

$$\begin{aligned} \mathrm{E}(\gamma) &= \frac{1 + \frac{m_1 m_2}{m_1 - m_2}(\zeta_1 - \zeta_2)}{1 + \frac{\delta_1 - \delta_2}{m_1 - m_2}} \\ Var(\gamma) &= \frac{\delta_1 - \delta_2}{m_1 - m_2}(\mathrm{E}(\gamma))^2 \\ n_X &= \frac{\delta_1}{\mathrm{E}(\gamma)}\left(\frac{c_{12}}{v_1} - \frac{\delta_2}{\delta_1}\right)\frac{\frac{m_1}{m_2} - 1}{\zeta_1 - \zeta_2} \\ n_Y &= \frac{\delta_2}{\mathrm{E}(\gamma)}\left(\frac{c_{12}}{v_2} - \frac{\delta_1}{\delta_2}\right)\frac{\frac{m_2}{m_1} - 1}{\zeta_2 - \zeta_1} \\ n_Z &= \frac{(m_1 - m_2)c_{12} + \frac{\delta_1 - \delta_2}{\zeta_1 - \zeta_2}}{m_1 m_2 \mathrm{E}(\gamma)(\zeta_2 - \zeta_1)}, \end{aligned} \tag{3}$$

   with all the denominators being non-zero.
2. *When* $n_X \neq n_Y$ *and* $Var(\gamma) = 0$ *(a purely binomial model with non-random* γ*), (2) has the unique*

solution

$$\gamma = 1 - \frac{v_1}{m_1} = 1 - \frac{v_2}{m_2},$$
$$n_Z = \frac{c_{12}}{\gamma(1-\gamma)}, \quad n_X = \frac{m_1}{\gamma} - n_Z, \quad n_Y = \frac{m_2}{\gamma} - n_Z. \tag{4}$$

3. *When $n_X = n_Y$, (2) has infinitely many solutions.*

It is not hard to see that a similar model with conditionally Poissonian distributions (instead of binomial distributions) is not identifiable. We summarise briefly other model candidates and their identifiability in our context in Table 1.

We close this section by discussing the use of the conditionally binomial model in more complex networks. As an example, consider a line topology with 4 nodes $\{1,2,3,4\}$, 3 directed links $\{\ell_1,\ell_2,\ell_3\} := \{(1,2),(2,3),(3,4)\}$ and 6 routes $\mathcal{R} := \{r_1,r_2,r_3,r_4,r_5,r_6,\} := \{\{\ell_1\},\{\ell_2\},\{\ell_3\},\{\ell_1,\ell_2\},\{\ell_2,\ell_3\},\{\ell_1,\ell_2,\ell_3\}\}$. Line topologies are relevant for vehicular traffic as well as for train passenger traffic.

Assume that, conditionally on a random parameter γ, the amount of vehicles on route r_i in a time period is a random variable X_i with distribution $Bin(n_i,\gamma)$, the X_is being independent given γ. Assume that the traffic O_i on link ℓ_i is measured, $i = 1,2,3$. Denote by A the 3×6 matrix (the route incidence matrix)

$$A_{ij} = 1_{\{\ell_i \in r_j\}}, \quad i = 1,2,3, \quad j = 1,\dots,6.$$

The counterpart of the moment Eqs. 2 can then be written in matrix form as

$$\mathrm{E}(O) = \mathrm{E}(\gamma) A\boldsymbol{n},$$
$$\mathrm{Cov}(O) = \mathrm{E}\left(\gamma(1-\gamma)\right) A\,\mathrm{diag}(\boldsymbol{n}) A^T + \mathrm{Var}(\gamma) A\boldsymbol{n}\boldsymbol{n}^T A^T, \tag{5}$$

where $O = (O_1,O_2,O_3)^T$ and $\boldsymbol{n} = (n_1,\dots,n_6)^T$. Since $\mathrm{Cov}(O)$ is symmetric, (5) amounts to 9 equations for the 8 unknowns $\mathrm{E}(\gamma)$, $\mathrm{Var}(\gamma)$, $n_1,\dots,n_6$. Substituting the first equation into the last term of the second, (5) is transformed into

$$\mathrm{E}(O) = \mathrm{E}(\gamma) A\mathbf{n},$$
$$\mathrm{Cov}(O) = \mathrm{E}\left(\gamma(1-\gamma)\right) A\,\mathrm{diag}(\boldsymbol{n}) A^T + \frac{\mathrm{Var}(\gamma)}{\mathrm{E}(\gamma)^2}\mathrm{E}(O)\mathrm{E}(O)^T. \tag{6}$$

In fact, Eq. 6 is valid for any network and route system with conditionally binomial traffic. The (i,j)-element of $A\,\mathrm{diag}(\boldsymbol{n})A^T$ is

$$[A\,\mathrm{diag}(\boldsymbol{n})A^T]_{ij} = \sum_{r_k\in\mathcal{R}:\ \ell_i\in r_k,\ \ell_j\in r_k} n_k. \tag{7}$$

It is straightforward to check (for example, by computing the determinant) that the linear map $(n_1,\dots,n_6) \mapsto A\,\mathrm{diag}(\boldsymbol{n})A^T$ is bijective. Thus, the vector $\boldsymbol{n}$ can be solved from the second equation of (6) as a linear combination of the elements of the matrix

$$\frac{1}{\mathrm{E}\left(\gamma(1-\gamma)\right)}\mathrm{Cov}(O) - \frac{\mathrm{Var}(\gamma)}{\mathrm{E}(\gamma)^2\mathrm{E}\left(\gamma(1-\gamma)\right)}\mathrm{E}(O)\mathrm{E}(O)^T.$$

Feeding this expression of $\boldsymbol{n}$ into the first equation of (6) yields three (effectively, two) non-linear equations for the two first moments of γ. Thus, the model is identifiable (except for some singular cases, cf. Proposition 1). An analytical solution of the equations is probably hard to find, but the optimisation approach of section 3.2 can be applied in practical cases. We point out that the above solution works in two steps separating the vector $\boldsymbol{n}$ and γ. This topic will be elaborated in section 4.3.

3 Traffic matrix estimation with the conditionally binomial model

Let us consider the estimation of the model parameters from the estimated first and second moments of the observations when $n_X \neq n_Y$ and $\mathrm{Var}(\gamma) > 0$.

3.1 Challenges of the use of proposition 1

Despite of the existence of the explicit solution (3), the model presents a principal challenge by depending on the accurate estimation of variances. To consider this in some detail, let $\gamma_1,\dots,\gamma_N$ be independent copies of a random variable γ with values in $(0,1)$, and for $k = 1,\dots,N$, let

Table 1 Summary of possible traffic models and their identifiablity

model	identifiability and comments
Poisson, extended Poisson	identifiable, not compatible with Tampere data, used in [12]
normal model (2-parameter models in general)	not identifiable; 6 parameters with 5 equations
normal model with common variance	not identifiable; unable to separate local and through traffic
conditionally Poissonian distributions, e.g., negative binomial distribution [14], Poisson-lognormal, Poisson - inverse Gaussian [18]	not identifiable; no bijection between the model parameters and the observation
conditionally binomial model	identifiable, has similar properties to Tampere data

U_k be an independent random variable with distribution $Bin(n, \gamma_k)$. The unbiased sample variance

$$s_N^2 = \frac{1}{N-1}\sum_{k=1}^{N}(U_k - \bar{U})^2, \quad \bar{U} = \frac{1}{N}\sum_{k=1}^{N} U,$$

has variance

$$\mathrm{Var}\left(s_N^2\right) = \frac{1}{N}\left(\mu_4 - \frac{N-3}{N-1}\sigma^4\right), \tag{8}$$

where $\sigma^2 = \mathrm{Var}(U)$ and $\mu_4 = \mathrm{E}\left((U - \mathrm{E}U)^4\right)$ (e.g., [8]).

With N large and n not too small, the standard deviation of s_N^2 is well approximated by

$$STD(s_N^2) \approx \sqrt{2}p(1-p)\frac{n}{\sqrt{N}} = \frac{\sqrt{2}(1-p)}{\sqrt{N}}\mathrm{E}(U). \tag{9}$$

Although the value of p can thus be consistently estimated as $p \approx 1 - s_N^2/\bar{U}$, the simultaneous estimation of both parameters of the binomial distribution is known to be difficult. [10] showed that there is no unbiased estimator of either parameter alone. The difficulty is intuitively obvious for the case of small p and large n, because $Bin(n, p)$ is then very close to $Poisson(np)$, a distribution with a single parameter. However, the usually powerful principle of maximum likelihood fails in this problem also for larger values of p, because the likelihood function turns out to be almost constant on the set $\left\{(n, p) : np = \bar{U}\right\}$.

When $\mathrm{Var}(\gamma) > 0$, an estimate of $STD(s_N^2)$ can be obtained by approximating U by a normal distribution with the same mean and variance. By Cochran's theorem [9], N i.i.d. samples from the normal distribution satisfy

$$(N-1)\frac{s_N^2}{\sigma^2} \sim \chi_{N-1}^2, \tag{10}$$

where χ_{N-1}^2 is the Chi-squared distribution with $N-1$ degreees of freedom. Figure 3 illustrates the confidence intervals of s_N^2/σ^2 at the risk levels $\alpha = 0.05, 0.1, 0.2$. Beyond $N = 10^4$, increasing the sample size decreases the uncertainty of the sample variance estimate extremely slowly.

The most frequently appearing element in the solution (3) is the difference $\zeta_1 - \zeta_2$ of the relative variances of O_1 and O_2. In order to provide satisfactory inference, the errors of the estimates ζ_1 and ζ_2 should be at the level of a fraction of their difference, say, at most $\xi|\zeta_1 - \zeta_2|$, with some not too high $\xi \in (0, 1)$.

Assume now that we have an i.i.d. N-sample from the conditionally binomial model, and let s_N^2 be the unbiased sample variance of O_1. The requirement $STD(\zeta_1) \le \xi|\zeta_1 - \zeta_2|$ is roughly equivalent to

$$\frac{STD(s_N^2)}{(\mathrm{E}(O_1))^2} \le \xi|\zeta_1 - \zeta_2| = \xi \cdot \frac{\mathrm{E}(\gamma) - \mathrm{E}\left(\gamma^2\right)}{(\mathrm{E}(\gamma))^2} \cdot \frac{|n_X - n_Y|}{(n_X + n_Z)(n_Y + n_Z)}. \tag{11}$$

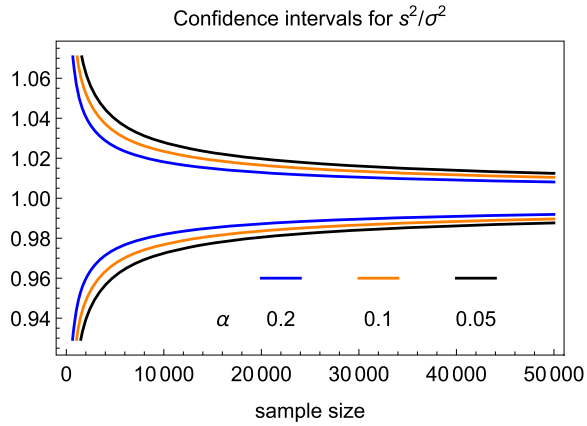

Fig. 3 Confidence intervals for the ratio s^2/σ^2 derived by using normal approximation for vehicle counts

Inserting to (10) $\sigma^2 = (n_X + n_Z)^2\mathrm{Var}(\gamma) + (n_X + n_Z)(\mathrm{E}(\gamma) - \mathrm{E}\left(\gamma^2\right))$ and $\mathrm{Var}\left(\chi_1^2\right) = 2$, writing

$$(n_X, n_Y, n_Z) = n(\beta_X, \beta_Y, \beta_Z)$$

with $\beta_X + \beta_Y + \beta_Z = 1$ and simplifying, we obtain from (11) the condition for N:

$$N \ge 2\left[\frac{\beta_Y + \beta_Z}{\xi|\beta_X - \beta_Y|}\left(1 + n\frac{\mathrm{Var}(\gamma)(\beta_X + \beta_Z)}{\mathrm{E}(\gamma) - \mathrm{E}\left(\gamma^2\right)}\right)\right]^2. \tag{12}$$

This expression reveals an important feature of the conditionally binomial model. When

$$n \le \frac{\mathrm{E}(\gamma) - \mathrm{E}\left(\gamma^2\right)}{\mathrm{Var}(\gamma)\,(\beta_X + \beta_Z)}, \tag{13}$$

the required number N of samples does not depend heavily on n. In the opposite case, however, the required N grows quadratically in n. To illustrate magnitudes, assume that $\mathrm{E}(\gamma) = 0.8$ and $STD(\gamma)$=0.05 — then $(\mathrm{E}(\gamma) - \mathrm{E}\left(\gamma^2\right))/\mathrm{Var}(\gamma) = 63$. Thus, the system size n affects the estimation precision adversely. Intuitively, the effect of the binomial fluctuations, on which the identification of n_Z is based, becomes with large n negligible in comparison to the effect of the variation of γ.

The estimation challenge is illustrated in Fig. 4. We simulated the model with γ being uniform on the interval $(0.6, 0.8)$ and the relative population sizes being $(\beta_X, \beta_Y, \beta_Z) = (2/6, 1/6, 3/6)$, and studied the accuracy of estimation from (3) with system sizes $n = 60$ and 600, which are below and above the critical magnitude given in (13). The model parameters were estimated in both cases in hundred samples of size $N = 5000$, computing the estimates from increasing subsamples to see the speed of convergence. Figure 4 presents the mean relative

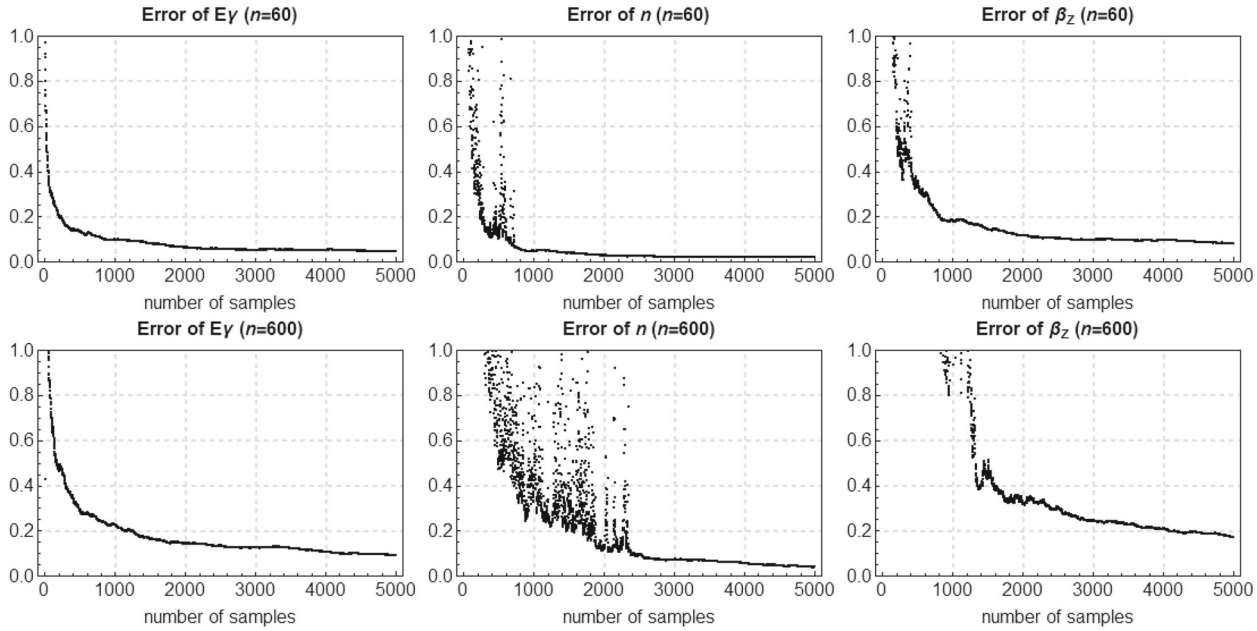

Fig. 4 Mean relative errors of estimates of $\mathrm{E}(\gamma)$ (left), n (middle) and β_Z (right), computed by (3) from hundred series of simulated observations. In each series, the estimates are calculated in an incremental fashion by increasing the sample size (given on the horizontal axis). The true parameter values were $\mathrm{E}(\gamma) = 0.7$, $\mathrm{Var}(\gamma) = 1/300$, $\beta_X = 2/6$, $\beta_Y = 1/6$, $\beta_Z = 3/6$, and $n = 60$ in the upper row and $n = 600$ in the lower row

errors of the estimates of $\mathrm{E}(\gamma)$, $n = n_X + n_Y + n_Z$ and $\beta_Z = n_Z/(n_X + n_Y + n_Z)$. The experiment suggests that the smaller system can be identified rather well with about 600 samples, whereas estimates from the larger system can easily be nonsense (e.g., negative) even when there are several thousands of samples (individual simulation runs were quite different, and the shapes of the point clouds varied). Note that the number of daily samples with some degree of homogeneity can hardly exceed 1000 in the real world.

3.2 Optimisation approach

We discuss the optimisation approach first in the context of the Minicity problem as it illustrates the main elements of the approach and serves our real traffic case study.

A natural approach to solve the traffic matrix estimation problem in both directions simultaneously is to minimise the squarred error of moment Eqs. 2 and $(2)^{EW}$. However, the covariance equations pose an additional challenge in the minimisation approach: if the last equation in (2) produces an error $\ell = c_{12} - \left((n_X + n_Z)(n_Y + n_Z)\mathrm{Var}(\gamma) + n_Z(\mathrm{E}(\gamma) - \mathrm{E}\left(\gamma^2\right)\right)$ with parameters $(n_X, n_Y, n_Z, \mathrm{E}(\gamma), \mathrm{Var}(\gamma))$, then the modified parameters $(n_X + \epsilon, n_Y + \epsilon, n_Z - \epsilon, \mathrm{E}(\gamma), \mathrm{Var}(\gamma))$, where $\epsilon = 2\ell/(\mathrm{E}(\gamma) - \mathrm{E}\left(\gamma^2\right))$, produce an equal error with an opposite sign. Thus the minimisation of the moment equations cannot uniquely determine optimal parameter values as the equations allow shifting traffic between local and through traffic.

To avoid this difficulty, we require that the covariance equations (the last equations of (2) and $(2)^{EW}$) be solved exactly. In the general form, the problem is

$$\min \sum_i \left(\bar{O}_i - \mathrm{E}\left(O_i\right)\right)^2 + \kappa \sum_i \left(s^2(O_i) - \mathrm{Var}\left(O_i\right)\right)^2$$
$$s.t.\ \mathrm{Cov}\left(O_i, O_j\right) = q_{ij} \quad \forall i, j,$$

where i, j run over traffic count measurement locations, $\bar{O}_i$ is the sample mean over the days, and s^2 and q_{ij} stand for sample variance and covariance, respectively. Expressions containing the parameters are $\mathrm{E}\left(O_i\right)$, $\mathrm{Var}\left(O_i\right)$, and $\mathrm{Cov}\left(O_i, O_j\right)$, given by (2) and its counterpart $(2)^{EW}$; κ is a weight and scaling parameter.

In our case, the solution of the quadratic equation for n_Z is

$$n_Z = \Big\{ -\mathrm{E}(\gamma) + (\mathrm{E}(\gamma))^2 - (n_X + n_Y - 1)\mathrm{Var}(\gamma) + \Big(\big[-\mathrm{E}(\gamma) + (\mathrm{E}(\gamma))^2 - (n_X + n_Y - 1)\mathrm{Var}(\gamma) \big]^2 + 4\mathrm{Var}(\gamma)(c_{12} - n_X n_Y \mathrm{Var}(\gamma)) \Big)^{1/2} \Big\} \Big/ (2\mathrm{Var}(\gamma)), \tag{14}$$

and a similar one holds for the East-West counterpart m_Z. The quadratic equation has exactly one positive solution, because $c_{12} - \mathrm{Var}(\gamma) n_X n_Y > 0$ by the positivity of n_Z. To summarise, in the bidirectional traffic matrix

estimation we search for $(n_X, n_Y, m_X, m_Y, \mathrm{E}(\gamma), \mathrm{Var}(\gamma))$ that minimise the cost function

$$\begin{aligned}
&(\mathrm{E}(O_1) - (n_X + n_Z)\mathrm{E}(\gamma))^2 \\
&+(\mathrm{E}(O_2) - (n_Y + n_Z)\mathrm{E}(\gamma))^2 \\
&+(\mathrm{E}(O_3) - (m_X + m_Z)\mathrm{E}(\gamma))^2 \\
&+(\mathrm{E}(O_4) - (m_Y + m_Z)\mathrm{E}(\gamma))^2 \\
+\kappa \times \Big\{ &(\mathrm{Var}(O_1) - (n_X + n_Z)^2\mathrm{Var}(\gamma) \\
&- (n_X + n_Z)(\mathrm{E}(\gamma) - \mathrm{E}(\gamma^2)))^2 \\
&+(\mathrm{Var}(O_2) - (n_Y + n_Z)^2\mathrm{Var}(\gamma) \\
&- (n_Y + n_Z)(\mathrm{E}(\gamma) - \mathrm{E}(\gamma^2)))^2 \\
&+(\mathrm{Var}(O_3) - (m_X + m_Z)^2\mathrm{Var}(\gamma) \\
&- (m_X + m_Z)(\mathrm{E}(\gamma) - \mathrm{E}(\gamma^2)))^2 \\
&+(\mathrm{Var}(O_4) - (m_Y + m_Z)^2\mathrm{Var}(\gamma) \\
&- (m_Y + m_Z)(\mathrm{E}(\gamma) - \mathrm{E}(\gamma^2)))^2 \Big\},
\end{aligned} \tag{15}$$

where n_Z is given by (14) and m_Z respectively. As the estimation of the first moments is more robust than the estimation of the second moments, we require that the first moments to be fitted more accurately than the second moments. This is achieved by the weight parameter κ that multiplies the fitting error of the second order moments, see [12] and [13]. Additionally, κ scales the different orders of magnitude in the first and second moments.

The optimisation approach of a more general topology follows from Eq. 6 in section 2.3. The first task is to solve the vector $\boldsymbol{n}$ as explained in that section. The optimisation is applied to the observed link means to solve for $\mathrm{E}(\gamma)$ and $\mathrm{Var}(\gamma)$. Note that this process naturally separates the parameters expressing the number of potential vehicles in each OD pair and the common intensity variable γ. The discussion of our case study in section 4.3 elaborates more the interplay of these parameters, as well as benefits of separating them.

4 Case study: morning traffic in the city of Tampere

4.1 Data

We studied the O-D estimation from a 15 minute interval traffic count data collected in the city of Tampere during 2011-2014. The city topology and the measurement points are illustrated in Fig. 5.

Unfortunately, various construction works caused interruptions in data and we selected the northern motorway with measurement locations Santalahti and Tampella for the traffic matrix estimation in the Minicity model. This road is a fast way to enter the city centre, but there is a large amount of traffic passing by the city centre.

An earlier study of the data suggested that the weekday morning traffic between 06:00 and 10:00 `A.M.` indicates the traffic flows most clearly. In the afternoons between 3:00 and 9:00 `P.M.` there can be some correlation between measurement points due to the traffic flows, but it is hard to detect from the general simultaneous activity in all the directions of the traffic. For other time periods, traffic bursts at the measurement locations near the city centre can be considered independent.

For the Minicity traffic matrix estimation, we select two morning periods 7:45 - 8:45 `A.M.` and 8:45 - 9:45 `A.M.` on weekdays from Monday through Thursday. The morning rush hour tends to shift slightly later on Fridays. Also, we exclude all public holidays, school holidays, isolated working days next to a public holiday, days between Christmas Eve and the New Year as well as months June – August due to lower traffic volumes. In this way, we end up with N=528 and N=522 days of traffic count measurements for the former and the latter rush hour period, respectively.

4.2 Estimation results

Although we removed atypical working days a priori there have been accidents and other events that produce outlier traffic counts that we removed by FAST-MCD covariance estimation [19]. We can safely assume that our data set contains less than 25% of contamination and configure the algorithm as recommended in [19]. To minimise the cost function of (15) with $\kappa = 0.00001$, a global optimisation routine `NMinimize` in Mathematica is utilised in algorithm autoselection mode. Also, the region of minimisation is constrained to be $(n_X, n_Y, m_X, m_Y, \mathrm{E}(\gamma), \mathrm{Var}(\gamma))$ such that $n_Z > 0, n_Y > 0, m_X > 0, m_Y > 0, \mathrm{E}(\gamma) \leq 1, \mathrm{Var}(\gamma) \geq 0, c_{12} - n_X n_Y \mathrm{Var}(\gamma) > 0, \mathrm{Var}(\gamma) \geq 0, c_{34} - m_X m_Y \mathrm{Var}(\gamma) > 0$. The two last constraints assure that n_Z and m_Z have positive solutions. The optimisation estimates the mean and the variance of the random variable γ. When plotting solutions against the measured data, we model γ by a beta distribution. The beta-binomial distribution has properties that fit well with our observations on the index of dispersion and the coefficient of the variation of the measured traffic, see section 2.2. However, some other distribution for γ may provide a better fit to the data.

4.2.1 *Morning traffic 7:45-8:45* `A.M.`

The minimum cost of Eq. 15 is 89.70 with the optimal parameter values provided at the top of the Table 2.

If we wish to model γ with a beta distribution, then the distribution parameters would be $\alpha = 43.54$ and $\beta = 4.59$.

The majority of vehicles travelling from West to East pass both measurement locations. This is expected, because the western section of Tampere has large housing districts (single houses), whereas the universities, hospitals, and other large offices are located in the eastern part

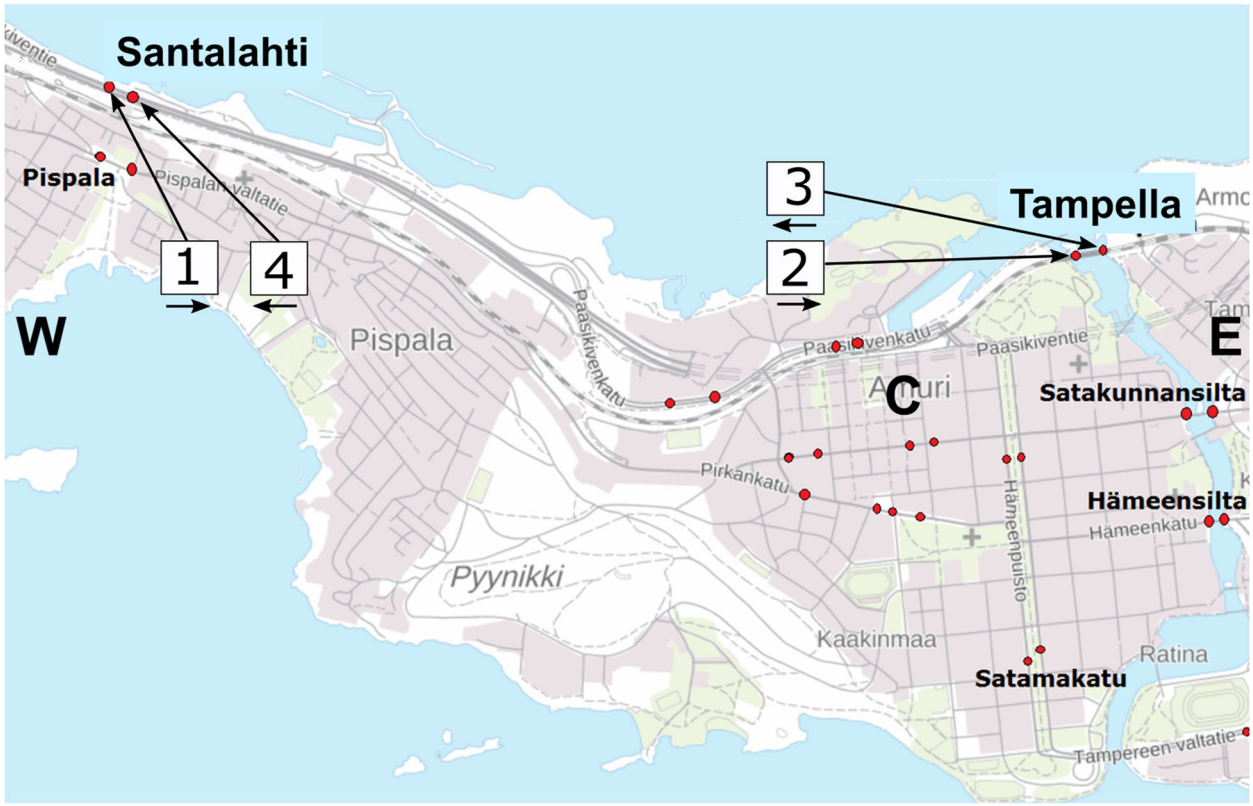

Fig. 5 Illustration of inductive loop measurement points in the city of Tampere. This research utilises locations Santalahti and Tampella and our measurement points are denoted as in Fig 1. The city centre, denoted by C, is located at southwest of Tampella

of Tampere. 622 potential vehicles arriving to the city centre is a feasible estimate. The value $\mathrm{E}(\gamma) = 0.91$ is somewhat high, but well in line with analysis of other rush hour traffic counts. As expected, in the opposite direction, most vehicles pass both measurement locations and only a few vehicles start from the city centre in order to reach the western part of the city.

The measured and estimated vehicle counts are illustrated by smooth histograms in Fig. 6, with an approximation by the above beta-binomial distribution. The traffic counts at the measurement locations O_1, O_2 and O_4 can be estimated reasonably well. There is peakness and skewness present in the observed traffic counts that is challenging to model with simple statistical models. The Kolmogorov – Smirnov distance between the measured data and the model, i.e, between the blue and red distributions, is 0.11, 0.08, 0.10, and 0.11 (at O_1, O_2, O_3, and O_4, respectively). By Kolmogorov – Smirnov test, the measured and modelled distributions are different at all the locations.

Table 2 Optimal parameter values for Monday - Thursday morning traffic

Optimal parameter values at 7:45-8:45 AM							
n_X	n_Y	n_Z	m_X	m_Y	m_Z	$\mathrm{E}(\gamma)$	$\mathrm{Var}(\gamma)$
622	49	2140	221	89	1273	0.91	0.0017
Optimal parameter values at 8:45-9:45 AM							
n_X	n_Y	n_Z	m_X	m_Y	m_Z	$\mathrm{E}(\gamma)$	$\mathrm{Var}(\gamma)$
783	170	2227	28	116	1782	0.54	0.0007

4.2.2 *Morning traffic 8:45-9:45* A.M.

For Monday-Thursday morning traffic at 8:45-9:45 A.M., the minimum cost of (15) is 1396, with the parameter values given at the bottom of Table 2. The beta distribution parameters for γ are $\alpha = 180.00$ and $\beta = 153.88$.

This period ends the Monday-Thursday morning rush hour traffic. Now the result is qualitatively different with a clearly lower value $\mathrm{E}(\gamma) = 0.54$. Illustration in Fig. 7 shows that the shape of the observed vehicle count distribution is more rounded and flatter, but skewness of the observed distribution is clearly visible at locations O_3 and O_4. The beta-binomial distribution, with estimated parameter values, is more symmetric. The Kolmogorov – Smirnov distance between the measured data and the model is 0.09, 0.14, 0.15, and 0.06 at O_1, O_2, O_3, and O_4, respectively. By Kolmogorov – Smirnov test, the

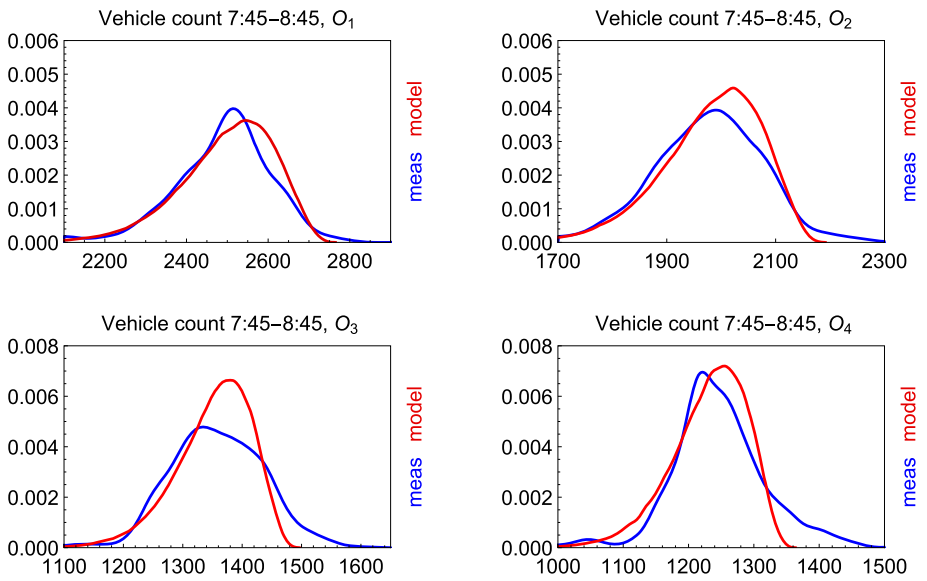

Fig. 6 Smooth histogram illustrations of measured (blue) and modelled (red) vehicle counts on Monday – Thursday mornings between 7:45 – 8:45 AM. A beta distribution has been taken to model γ

measured and modelled distributions are different at all the locations, but at O_4 the p-value is 0.047.

4.3 Robustness of the estimation results

We examine the robustness of the estimation results in two ways. First, we apply resampling to the original data, and then we examine the optimal parameter values with generated data.

Resampling: The resampling data are generated by selecting randomly without replacement 90% of the Monday – Thursday morning data to produce a collection of 1000 estimation samples. For morning traffic between 7:45 – 8:45 A.M. , we study closely 162 estimation results with good fit to the data. In these results, the value of $\mathrm{E}(\gamma)$ varies from 0.55 to 0.95, i.e., the probability of executing a journey could also be clearly lower than the one estimated

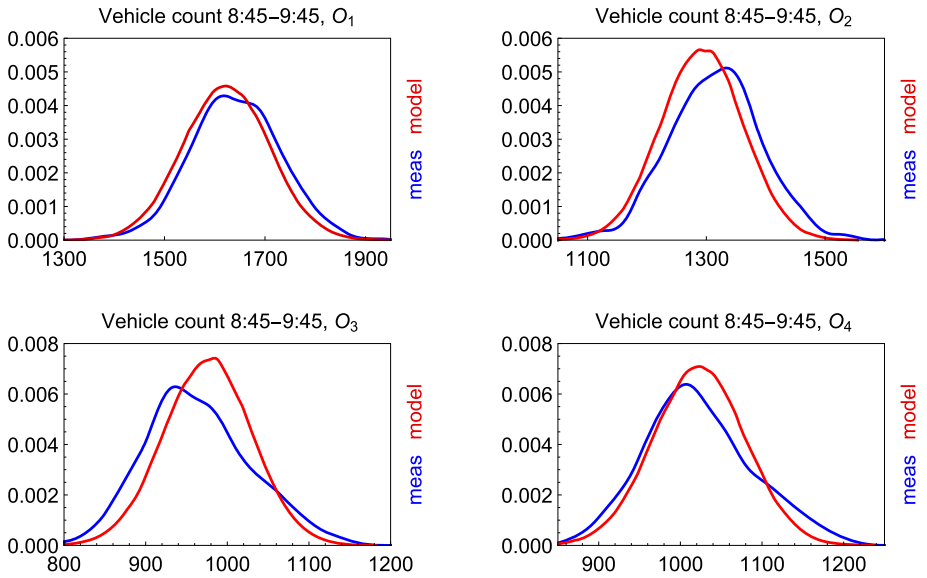

Fig. 7 Smooth histogram illustrations of traffic between 8:45 - 9:45 AM, notation as in Figure 6

from the original data. Clustering analysis of the results indicates that most solutions are around the original data estimate, but there could be an alternative model candidate that has a lower $\mathrm{E}(\gamma) \approx 0.65$. We note that this value of $\mathrm{E}(\gamma)$ would be closer to the results of the traffic matrix estimation for 8:45 – 9:45 `A.M.`. The traffic matrix probably evolves during the morning hours. For the peak traffic, the value of $\mathrm{E}(\gamma)$ is high, and when the traffic volumes decrease, also the value of $\mathrm{E}(\gamma)$ decreases.

The same processes are repeated with the morning traffic data for 8:45 – 9:45 `A.M.`, where 201 good estimation results that are studied in more detail. These results are in line with the ones from the original data.

Generated data: Next, we generate from a beta-binomial distribution 1000 samples each of size 500, with parameters that resemble the Tampere city case, and examine the accuracy of our traffic matrix estimation technique. Note that this realistic experiment is a challenging one; the parameters are not in the range of efficient inference identified in section 3.1, because the m_X and m_Y values for East–West local traffic are close to each other. Results of section 3.1 and parameters at the top line of Table 2 allow to calculate lower bounds for the required number of samples: in the West–East direction, Eq. 12 with $\xi = 1$ gives rise to a lower bound of 100 119 samples. In comparison, in the East–West direction, the corresponding lower bound is 213 814 samples, which reflects the difficulty of the estimation when m_X and m_Y values are close to each other. Eq. 13 provides an upper bound for the total number of vehicles per traffic direction, $n_X + n_Y + n_Z$ or $m_X + m_Y + m_Z$, so that the required number of samples does not grow quadratically in the total number of vehicles. These upper bounds are around 50 vehicles, whereas the estimated values for the first morning period give rise to 2811 vehicles in West-East direction and 1583 in the opposite direction. Thus this experiment is extremely challenging for the optimisation.

In the estimation experiment with simulated samples and known parameter values, the true value of the West–East ratio $n_X/(n_X + n_Z)$ is captured very well, but the East–West ratio $m_X/(m_X + m_Z)$ tends to be estimated too low. Further, the estimation tends to favour high n or m values together with low $\mathrm{E}(\gamma)$ values. In the model, the errors in the n and m parameters versus $\mathrm{E}(\gamma)$ compensate each other, which explains the good fits to the generated traffic observations. We conclude that the sample size of 500 is not large enough to estimate the exact parameters values in a robust and reliable manner.

In practice, there needs to be additional information to judge the correct levels of the n or m parameters and $\mathrm{E}(\gamma)$. There is freedom to increase one and to decrease the other while keeping the product intact, which is the basic challenge of $Bin(n, p)$ estimation with both p and n unknown. [10] note that the estimation of n and p are linked together. On the other hand, in our model the variable γ, common to all flows, limits the degree of freedom. Our further experiments indicated that the balancing the overestimation/underestimation of the n or m values with $\mathrm{E}(\gamma)$ vanishes when some parameters are fixed to their true values.

5 Conclusions and future perspectives

The analysis of daily quarter-hour traffic count time series data on city traffic in Tampere shows strong positive correlations even between measurement points that share no traffic flows. This suggests the consideration of doubly stochastic traffic models, in which some common factor influences the traffic volumes of all O-D flows. Because similar positive correlatedness of vehicular traffic can be expected to hold rather generally, it deserves more attention as a challenge for modelling and statistical inference. In this paper, we studied the ability of a conditionally binomial model to utilise the correlations caused by shared traffic, despite of the additional correlations caused by the common activity factor. The model benefits from correct statistical properties as well as a rather intuitive role of parameters.

We focused on a simple network model, but we also showed that conditionally binomial traffic models are identifiable in rather general network and route scenarios. Our study at the end of section 2.3 indicates that optimisation might be the best method for practical solutions, but a detailed analysis of this matter remains an open research topic.

We examined solving the O-D matrix problem using the first and second order statistics of the observed link counts. The conditionally binomial model can be solved exactly and the solution is numerically feasible when traffic volumes are sufficiently small. The analysis reveals parameter regions in which estimation challenges are expected. Unfortunately, our real traffic case of the city of Tampere falls in such a parameter region. However, approximate solutions for the O-D matrix could be obtained by optimisation methods. Our accuracy studies indicated that the solutions may suffer from simultaneous over- and underestimation of parameters, similar to the well known challenge of estimating $Bin(n, p)$ when both n and p are unknown. We regard link counts as the primary source of information, but acknowledge, similarly to [15], the benefits of some additional information. The context of vehicular traffic allows incorporating additional geoinformatics, see e.g. [22] and separating the n, m-variables from γ. We see those as succesfull future approaches.

Abbreviations

O-D: origin– destination; IDC: index of dispersion; CoV: coefficient of variation; $Bin(n, p)$: binomial distribution; STD: standard deviation; n, m: population size parameters; N: number of samples

Acknowledgements
We thank Kimmo Ylisiurunen and Aleksi Vesanto from InfoTripla Oy for informative discussions as well as for the preparation of the traffic data and permission to publish the data. We thank Dario Gasbarra for discussions on statistical methods and Fanny Malin for preparing Figure 5.

Authors' contributions
All authors have contributed ta all parts af this work. All authors read and approved the final manuscript.

Funding
This research work was funded by the Academy af Finland project 294763 Stomograph. Refining resuls and publishing was funded by EU ECSEL project 737494 MegaMart2 and Business Finland project 123137 RAGE, and VTT's internal funding.

Availability of data and materials
The datasets analysed during the current study are available in the CERN's Zenodo repository, https://doi.org/10.5281/zenodo.2359339.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Techical Research Centre af Finland, VTT Ltd., P.O. Box 1000, 02044 VTT, Espoo, Finland. [2]University af Helsinki, Department af Mathematics and Statistics, P.O. Box 64, FI-00014 Helsinki, Finland.

Received: 7 January 2020 Accepted: 26 May 2020
Published online: 17 June 2020

References

1. Bera, S, & Rao, KVK (2011). Estimation of origin-destination matrix from traffic counts: the state of the art. *European Transport*, *49*, 3–23.
2. Bifulco, GN, Cartenì, A, Papola, A (2010). An activity-based approach for complex travel behaviour modelling. *European Transport Research Review*, *2*(4), 209–221.
3. Castillo, E, Jiménez, P, Menéndez, JM, Conejo, AJ (2008). The observability problem in traffic models: Algebraic and topological methods. *Trans Intell Transport Sys*, *9*(2), 275–287.
4. Castillo, E, Calviñ, A, Nogal, M, Lo, HK (2014). On the probabilistic and physical consistency of traffic random variables and models. *Computer-Aided Civil and Infrastructure Engineering*, *29*(7), 496–517.
5. Castillo, E, Conejo, AJ, Menéndez, JM, Jiménez, P (2008). The observability problem in traffic network models. *Computer-Aided Civil and Infrastructure Engineering*, *23*(3), 208–222.
6. Castillo, E, Grande, Z, Calviñ, A, Szeto, WY, Lo, HK (2015). A state-of-the-art review of the sensor location, flow observability, estimation, and prediction problems in traffic networks. *Journal of Sensors*, *26*. https://doi.org/10.1155/2015/903563.
7. Castillo, E, Menéndez, JM, Sánchez-Cambronero, SS, Calviño, A, Sarabia, JM (2014). A hierarchical optimization problem: Estimating traffic flow using gamma random variables in a Bayesian context. *Computers & Operations Research*, *41*, 240–251. https://doi.org/10.1016/j.cor.2012.04.011.
8. Cho, E, Cho, MJ, Eltinge, JE (2005). The variance of sample variance from a finite population. *International Journal of Pure and Applied Mathematics*, *21*(3), 387–394.
9. Cochran, W (1934). The distribution of quadratic forms in a normal system, with applications to the analysis of covariance. *Mathematical Proceedings of the Cambridge Philosophical Society*, *30*(2), 178–191. https://doi.org/10.1017/S0305004100016595.
10. DasGupta, A, & Rubin, H (2005). Estimation of binomial parameters when both n, p are unknown. *Journal of Statistical Planning and Inference*, *130*(1), 391–404. https://doi.org/10.1016/j.jspi.2004.02.019.
11. Frederix, R, Viti, F, Tampére, CMJ (2013). Dynamic origin–destination estimation in congested networks: theoretical findings and implications in practice. *Transportmetrica A: Transport Science*, *9*(6), 494–513.
12. Hazelton, ML (2003). Some comments on origin-destination matrix estimation. *Transportation Research Part A: Policy and Practice*, *37*(10), 811–822.
13. Hazelton, ML (2008). Statistical inference for time varying origin-destination matrices. *Transportation Research Part B: Methodological*, *42*(6), 542–552.
14. Hazelton, ML (2015). Network tomography for integer-valued traffic. *Annals of Applied Statistics*, *9*(1), 474–506. https://projecteuclid.org/euclid.aoas/1430226101.
15. Hazelton, ML, & Parry, K (2016). Statistical methods for comparison of day-to-day traffic models. *Transportation Research Part B: Methodological*, *92*, 22–34. https://doi.org/10.1016/j.trb.2015.08.005.
16. Kilpi, J, Norros, I, Kuusela, P, Malin, F, Räty, T (2020). Robust methods and conditional expectations for vehicular traffic count analysis. *European Transport Research Review*, *12*(1). https://eur03.safelinks.protection.outlook.com/?url=https%3A%2F%2Fdoi.org%2F10.1186%2Fs12544-020-0399-8&data=02%7C01%7CPirkko.Kuusela%40vtt.fi%7C984085732b454e37eb9008d80732c560%7C68d6b592500843b59b0423bec4e86cf7%7C0%7C0%7C637267261928502462&sdata=6cnlCbGMwDd0h1x6gSj%2B5R08SQ2%2FOJSl3A24bP52f5g%3D&reserved=0 https://doi.org/10.1186/s12544-020-0399-8.
17. Parry, K, Watling, DP, Hazelton, ML (2016). A new class of doubly stochastic day-to- day dynamic traffic assignment models. *EURO Journal on Transportation and Logistics*, *5*(1), 5–23.
18. Perrakis, K, Karlis, D, Cools, M, Janssens, D (2015). Bayesian inference for transportation origin–destination matrices: the poisson–inverse gaussian and other poisson mixtures. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *178*(1), 271–296.
19. Rousseeuw, PJ, & Van Driessen, K (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics*, *41*(3), 212–223. https://doi.org/10.1080/00401706.1999.10485670.
20. Singhal, H, & Michailidis, G (2007). Identifiability of flow distributions from link measurements with applications to computer networks. *Inverse Problems*, *23*(5), 1821–1849.
21. Vardi, Y (1996). Network tomography: Estimating source-destination traffic intensities from link data. *Journal of the American Statistical Association*, *91*(433), 365–377.
22. Wang, S, Yu, D, Ma, X, Xing, X (2018). Analyzing urban traffic demand distribution and the correlation between traffic flow and the built environment based on detector data and POIs. *European Transport Research Review*, *10*(2), 50.
23. Yudi Yang, Y, & Yueyue Fan, Y (2015). Data dependent input control for origin–destination demand estimation using observability analysis. *Transportation Research Part B: Methodological*, *78*(C), 385–403.
24. Yang, Y, Fan, Y, Royset, JO (2019). Estimating probability distributions of travel demand on a congested network. *Transportation Research Part B: Methodological*, *122*, 265–286.
25. Yang, Y, Fan, Y, Wets, RJB (2018). Stochastic travel demand estimation: Improving network identifiability using multi-day observation sets. *Transportation Research Part B: Methodological*, *107*, 192–211.

Publisher's Note